DOCUMENT RESUME

ED 268 139 TM 850 778

AUTHOR Lord, Frederic M.; Wild, Cheryl L.
TITLE Contribution of Verbal Item Types in the GRE General Test to Accuracy of Measurement of the Verbal Scores.
INSTITUTION Educational Testing Service, Princeton, N.J.
SPONS AGENCY Graduate Record Examinations Board, Princeton, N.J.
REPORT NO ETS-RR-85-29; GREB-PR-84-6P
PUB DATE Sep 85
NOTE 23p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *College Entrance Examinations; *Error of Measurement; Graduate Study; Higher Education; *Item Analysis; Latent Trait Theory; Reading Comprehension; Reading Tests; Scaling; *Scores; Statistical Studies; *Test Items; Test Validity; *Verbal Tests
IDENTIFIERS Analogy Test Items; Antonyms; *Graduate Record Examinations; Sentence Completion Test

ABSTRACT

This study compares the contribution to measurement accuracy of the verbal score of each of four verbal item types included in the Graduate Record Examinations (GRE) General Test. Comparisons are based on item response theory, a methodology that allows the researcher to look at the accuracy of individual points on the score scale. This methodology is based on the assumption that the four verbal item types measure the same verbal ability. Since the results of the study do indicate that the reading comprehension item type measures something slightly different from what is measured by sentence completion, analogy, or antonym item types, only tentative conclusions may be drawn. The antonym item type contributes the most accuracy of the four item types for scores above about 550. Analogy items contribute to the measurement accuracy of verbal ability throughout the score range. This is especially true when item types are matched on verbal difficulty. These results suggest that the analogy and antonym item types are useful for maintaining accuracy of the verbal score scale at the upper levels. Eliminating these items might have a serious impact on the validity of the GRE verbal score in the upper regions of the scale. Studies of the validity of item types at the upper score range using external criteria would be necessary to understand the exact contribution of the item types to the validity of the test. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED268139

GRE

GRADUATE RECORD EXAMINATIONS

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

H. Weickenmiller

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 850 478

CONTRIBUTION OF VERBAL ITEM TYPES IN THE GRE GENERAL TEST TO ACCURACY OF MEASUREMENT OF THE VERBAL SCORES

Frederic M. Lord

and

Cheryl L. Wild

GRE Board Professional Report GREB No. 84-6P
ETS Research Report 85-29

September 1985

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

EDUCATIONAL TESTING SERVICE, PRINCETON, NJ

Contribution of Verbal Item Types in the GRE General Test to Accuracy of Measurement of the Verbal Scores

Frederic M. Lord

and

Cheryl L. Wild

GRE Board Professional Report No. 84-6P

September 1985

Copyright © 1985 by Educational Testing Service
All rights reserved

CONTRIBUTION OF VERBAL ITEM TYPES IN THE GRE GENERAL TEST TO ACCURACY OF MEASUREMENT OF THE VERBAL SCORES

Abstract

The main purpose of this study is to compare the contribution to measurement accuracy of the verbal score of each of the four verbal item types included in the GRE General Test. Comparisons are based on item response theory, a methodology that allows the researcher to look at the accuracy of individual points on the score scale. This methodology is based on the assumption that the four verbal item types measure the same verbal ability. Since the results of the study do indicate that the reading comprehension item type measures something slightly different from what is measured by sentence completion, analogy, or antonym item types, only tentative conclusions may be drawn.

The antonym item type contributes the most accuracy of the four item types for scores above about 550. Analogy items contribute to the measurement accuracy of verbal ability throughout the score range. This is especially true when item types are matched on verbal difficulty. These results suggest that the analogy and antonym item types are useful for maintaining accuracy of the verbal score scale at the upper levels. Eliminating these items might have a serious impact on the validity of the GRE verbal score in the upper regions of the scale. Studies of the validity of item types at the upper score range using external criteria would be necessary to understand the exact contribution of the item types to the validity of the test.

CONTRIBUTION OF VERBAL ITEM TYPES IN THE GRE GENERAL TEST TO ACCURACY OF MEASUREMENT OF THE VERBAL SCORES

1. Purpose

The main objective of this study is to compare the contribution to measurement accuracy of each of the four verbal item types included in the Graduate Record Examinations (GRE) General Test verbal score. Comparisons were based on item response theory, a methodology that allows the researcher to look at the accuracy at individual points on the score scale. Information on the contribution of each item type can lead to test specifications that maximize test accuracy at various score points.

2. Background

The GRE verbal test has proven valuable as a selection criterion for graduate school for a number of years. Summaries of validity data show that the average correlation of the GRE verbal measure with graduate grade point average is higher than the average correlation of the quantitative score with graduate grade point average for humanities, social science, and biological science departments (Burton & Turner, 1983). The verbal score is also highly reliable (median of .92 for eight recent test editions).

Despite these positive characteristics, some recent research (Ramist, 1981) on the Scholastic Aptitude Test (SAT) and more recently on the Graduate Record Examinations (GRE) General Test (Wilson, 1985) calls into question the appropriateness of the current balance of item types. Wilson found that correlations of a subscore of GRE reading comprehension and sentence completion items were comparable to the correlation of the total verbal score (made up of antonym, analogy, reading comprehension, and sentence completion item types) with self-reported undergraduate grade point average.

Ramist (1981) found similar results for the SAT in a predictive validity model. Schrader has completed three studies (1984) that have looked in more depth at the SAT finding. Schrader first reviewed item and test analysis results and found that the analogy and antonym item types provided a substantially greater number of difficult items than reading comprehension or sentence completion item types. Although the fact that subscores were not based on separately timed sections complicated the interpretation of the results, it also appeared that antonym and analogy items were more likely to have lower biserial correlations with the total verbal score than reading comprehension and sentence completion items.

One result of these differences in difficulty and biserial correlations may be a lower correlation of a vocabulary subscore than a reading comprehension subscore with an external criterion. If the item difficulty and discrimination parameters were the same for each subscore, the subscores

may be equally valid. This hypothesis was supported by Schrader's third study where he selected subsets of items matched in terms of difficulty and found correlations between scores on these subsets and first-year college grades. Results indicated that the differences in validity of the four item types were relatively small. When subsets of items matched on difficulty were compared, reading comprehension had somewhat lower validity than the other three item types. However, these results were tentative because only one form of the SAT verbal was studied and because matched subsets included only eight items.

Validity of a test with respect to a criterion is limited by the index of reliability (Lord & Novick, 1968). All things being equal, a test with all questions of middle difficulty for the group of examinees will be most reliable. This type of test will have high accuracy in the middle of the score distribution, but little accuracy (and therefore low validity) at the extremes. A test with a wider distribution of item difficulties will have relatively greater accuracy at the extremes, but lower overall reliability, compared to the previously described test. Traditional indexes of reliability and validity, as averages over the total score range, give greater weight to the center of the distribution, the area of most frequent observations. However, these overall indexes may mask the increases in accuracy (and therefore, potentially, validity) at the extremes of the distribution.

If the assumptions of item response theory are valid, it is possible to use this theory to compute the accuracy of measurement at various score points. Although accuracy of measurement is a requirement for predictive validity of scores, validity does not necessarily follow from accuracy.

3. Criteria

Research cited above has suggested that a score on GRE reading items may be as good a predictor of graduate grade point average (GPA) as a score on the entire GRE verbal test. In reply, it has been urged that whereas a reading score is good for predicting GPA over middle ranges of ability, it is probably much less effective at high ability levels. This seems likely because reading items typically are not of sufficient difficulty to discriminate well among high-ability examinees. Attempts to write more difficult reading items are not as successful as attempts to write difficult antonym or analogy items.

A clear resolution of the relative merits of reading items and of verbal nonreading items would require validity studies on a large number of high-ability students at a variety of graduate schools, all of which assign course grades on a scale having the same meaning for all schools. In the absence of such data, the present study attempts to throw some light on the situation by substituting a different criterion for GPA. Because of this substitution, it will not be possible to draw rigorous conclusions about validity for predicting GPA.

The main criterion to be used here is number-right true score on the GRE verbal test. In section 7, number-right true score on GRE reading items is used instead. Results using the two criteria would be the same if the verbal measure were truly unidimensional.

4. Data

For the data used in this study, each 76-item GRE verbal test was administered together with a separately timed 38-item 'anchor test'.[1] The tests analyzed here are shown in Table 1. Since the anchor tests were built with pretested items to be parallel to the regular test editions, all 114 items from each administration are treated as representing typical GRE verbal test items. The four regular GRE editions plus the five anchor tests provide a total of 143 reading, 91 sentence completion, 143 antonym, and 117 verbal analogy items. All items are five choice. Only examinees who stated that their native language is English are used in the analyses.

Table 1

GRE Verbal Test Data Analyzed

Admin. Date	GRE Form	Anchor Test	No. of Examinees Total	No. of Examinees Native English Speakers*
2/82	3DGR2	D82	3,518	2,790
2/82	3DGR2	E17	3,380	2,655
4/82	3EGR1	E17	3,456	2,718
4/82	3EGR1	E20	3,396	2,681
10/82	3EGR4	E20	3,979	2,974
10/82	3EGR4	E85	3,890	2,899
2/83	3EGR2	E85	3,629	2,395
2/83	3EGR2	F1	3,474	2,248
			28,722	21,360

*Examinees who did not answer the question about their native language were omitted.

[1]The GRE General Test editions used in this study were from a large equating using anchor test equating as one methodology (GRE Board Research Project No. 81-16, in progress). The GRE General Test is currently equated using spiraling rather than anchor tests.

5. Method

Below the diagonal, Table 2 gives the intercorrelations between number-right score on the four different item types. The diagonal gives the Kuder-Richardson formula-20 reliabilities (alpha coefficients). Above the diagonal are the correlations corrected for attenuation. The data for this table were obtained from an older edition of the GRE verbal measure, 3CGR1. The editions in Table 1 are scored number right; the scores on 3CGR1 were number-right minus one-fourth number wrong. Table 2 would be most useful for evaluating the interrelations of the four item types if it were based on the same test forms used in Table 1. Unfortunately, this information is not obtained routinely. The correlations based on the formula scored test should provide an adequate estimate of the intercorrelations for these purposes.

Table 2

Correlations, Reliabilities (in diagonal), and Disattenuated Correlations for GRE Verbal Subtests, Form 3CGR1

	Reading Comprehension	Sentence Completion	Antonym	Analogy
Reading Comprehension	(.790)	.899*	.768*	.847*
Sentence Completion	.677	(.718)	.863*	.894*
Antonym	.632	.677	(.858)	.909*
Analogy	.649	.653	.726	(.743)

*Corrected for attenuation

In Table 2, the correlations corrected for attenuation are less than 1.0, indicating that different item types measure different traits. The correlations shown in Table 2 are lower than those found in the SAT. The likely reason is that GRE examinees have a smaller range of ability than the SAT examinees.

Since most correlations above the diagonal in Table 2 are less than .90, there are some doubts as to the appropriateness of unidimensional item response theory.[2] The study was carried out in spite of these doubts. To the extent that the unidimensionality assumption is only approximately correct, the results of the study must also be considered approximate.

All the responses given by native English-speaking examinees (see the last column of Table 1) were used in a single initial run of the computer program LOGIST (Wingersky, 1983) to obtain parameter estimates for all 494 items. It is assumed throughout this report that the probability of a correct answer to an item for examinees at a given ability level follows the three-parameter logistic model (Lord, 1980). This model is not standardly used for equating or item selection for the GRE General Test.

At any desired true-score level, the measurement effectiveness of the observed scores on two tests or subtests measuring the same dimension can be compared by means of the index of relative efficiency (R.E.). This index is given by the ratio

$$R.E.(y,x) = \frac{L^2\{x,\tau\}}{L^2\{y,\tau\}}$$

where τ is the number-right true-score level at which measurement effectiveness is to be evaluated, x and y are the observed scores on the two tests to be compared, $L\{x,\tau\}$ is the length of the confidence interval for estimating the examinee's ability τ from his or her observed score x, and $L\{y,\tau\}$ likewise for y. R.E. is readily computed from estimated

---

[2]Kingston and Dorans (1982) reviewed the feasibility of using item response theory for equating the GRE General Test when the test was administered using formula scoring instructions. They examined the local independence assumption in detail (in the unidimensional model case, this assumption is equivalent to an assumption of unidimensionality). They found that although the local independence assumption was violated, the three-parameter logistic model replicated observed verbal item data well. This was substantiated by reasonable results from item response theory true score equating. In the equating study currently in progress (GRE Board Research Project No. 81-16) using the current General Test with rights only scoring, the three-parameter logistic model did not replicate the verbal item data as well as it had under formula-scoring conditions. However, item response theory equating of the verbal measure appears to be quite accurate.

item parameters using standard formulas for τ and for $L\{\ ,\tau\}$ (Lord, 1980). For purposes of the next section, the 'verbal ability' criterion τ is the expected value of number-right observed score on all 494 items.

For the purposes of Section 7, a LOGIST run was made just on the responses to the 143 reading items. The 'reading ability' criterion τ used in Section 7 is expected number-right observed score on just the 143 reading items. This LOGIST run yielded an estimated 'reading ability' parameter for each of the 21,360 examinees. A second LOGIST run was then made, holding these 'reading ability' parameters fixed while estimating the item parameters for all 494 items. If the 494-item test is truly unidimensional, the R.E. curves computed from these item parameters will be the same as those found in Section 6 except for sampling fluctuations. Otherwise the curves will provide a rough indication of how well 'reading ability' can be estimated from observed scores on other types of verbal items.

6. Results for 'Verbal Ability'

Separately for each item type, Figure 1 shows how well the number-right score on each item type measures 'verbal ability'. More specifically, it shows the relative efficiency of number-right observed score on the item type compared to number-right observed score on the total GRE verbal test. This relative efficiency is a function of the true ability level, shown on the horizontal axis. The curves are adjusted for test length so that they do not reflect the number of items used for computing each score. The base line represents 'verbal ability' expressed numerically in terms of the usual GRE score scale.

The meaning of the figure can be understood from the following selected interpretations:

1. For examinees at a (true) 'verbal ability' level of 515, number-right observed score on 76 typical GRE antonym items estimates the examinee's 'verbal ability' with the same accuracy (same length confidence interval) as does the usual number-right observed score on the 76-item total GRE verbal test.

This conclusion and others listed below depend on the assumption that all item types measure the same 'verbal ability'. If this assumption is only an approximation, then the conclusions are only approximately true.

2. Item for item, number-right observed score on typical GRE reading items measures 'verbal ability' more accurately over the ability range from 225 to 495 than does the usual total GRE verbal test score. Reading items measure 'verbal ability' less accurately below 225 and above 495.

3. Compared to total GRE verbal test score, sentence completion number-right observed score has a relative efficiency of 1.6 for measuring the 'verbal ability' of examinees whose true 'verbal ability' is in the range 400 to 450. This means that the confidence intervals for estimating the ability of such examinees from their number-right observed scores on n typical total verbal items are $\sqrt{1.6} = 1.26$ times as long as the confidence intervals from n typical sentence completion items. Another, and simpler, way to say this is that the total verbal test would have to be lengthened by a factor of 1.6 (to 1.6 x 76 = 122 items) for it to measure as well as a 76-item sentence-completion test in the range from 400 to 450.

4. At high ability levels, the measuring of 'verbal ability' as defined by true score on the present test is best accomplished by the antonym items and to a lesser extent by the analogy items. In the range from 300 to 500, the sentence completion and the reading items are best. The effectiveness of the analogy items is roughly equal across all ability levels.

The GRE verbal test contains 22 reading, 14 sentence completion, 18 analogy, and 22 antonym items. Figure 2 shows the actual contribution of different item types to the existing total verbal test. Unlike Figure 1, Figure 2 has no adjustment for the number of items of each type.

The bottom curve in Figure 2 shows the relative efficiency of the 22 reading items compared to the 76-item total verbal test. The efficiency is low because the reading score is based on many fewer items than the total score. The middle curve in Figure 2 shows the relative efficiency of reading and sentence completion items combined. Thus the area between the reading curve and the combined curve represents the contribution of the sentence completion items to the measurement efficiency of the total test. Similarly, the area between the middle and the top curve represents the contribution of the analogy items. The area between the top curve and the horizontal line (representing a relative efficiency of 1.0) represents the contribution of antonym items. Clearly, the antonym items are the ones that in practice contribute the most measurement accuracy for higher ability examinees.

Figure 2 differs in another important way from Figure 1. The contribution of a test or subtest was computed not for number-right score as in Figure 1, but for an optimally weighted sum of item scores (0 or 1). The optimal weights are chosen to minimize the sampling error of the examinee's estimated true score. Optimal scores were used because the relative efficiencies of the separate subtests (item types) add up to exactly 1.0, allowing the measurement effectiveness to be partitioned into additive contributions made by the separate item types.

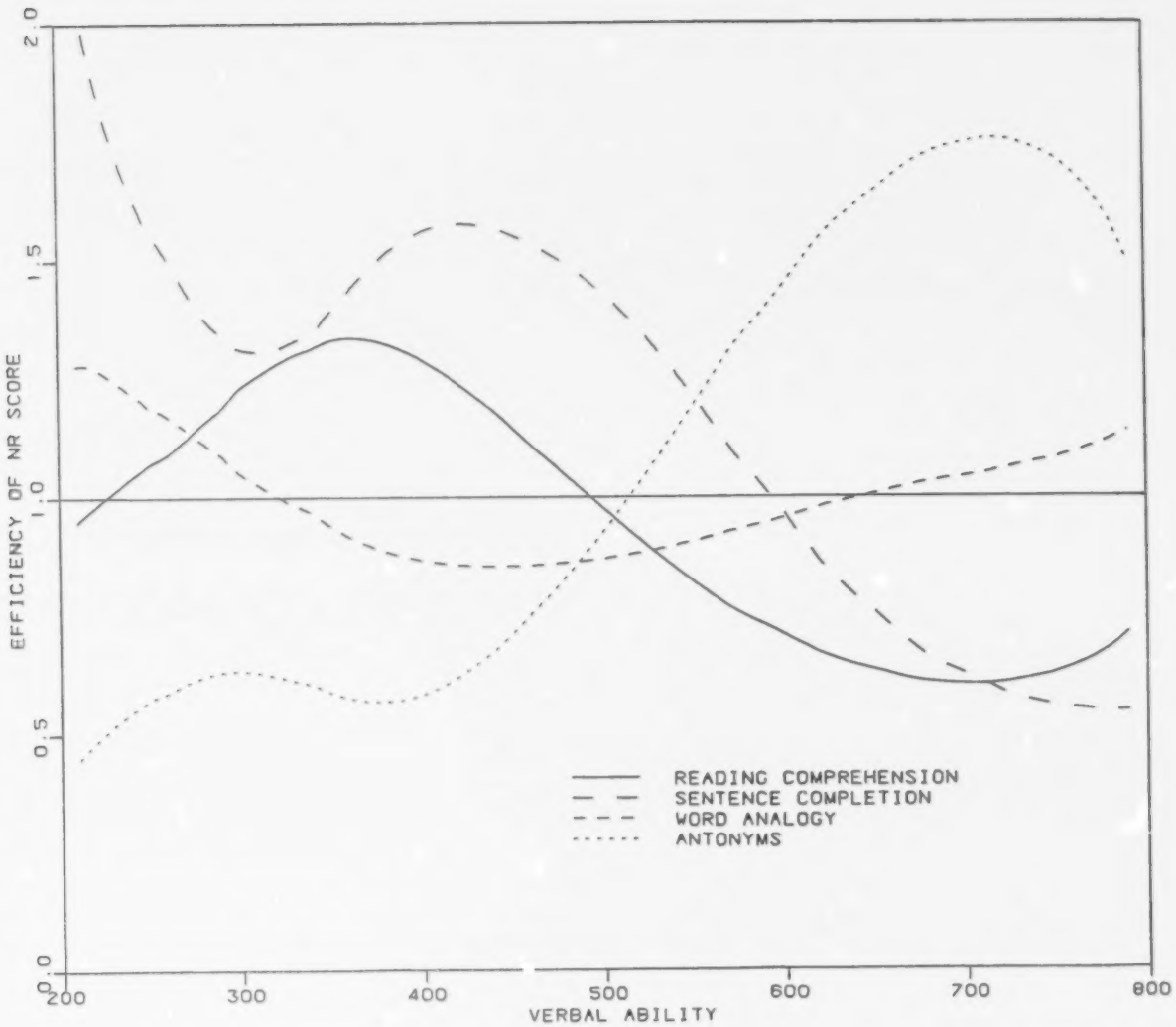

Figure 1. Efficiency, item for item, of four item types, relative to the total GRE verbal test, for measuring 'verbal ability'.

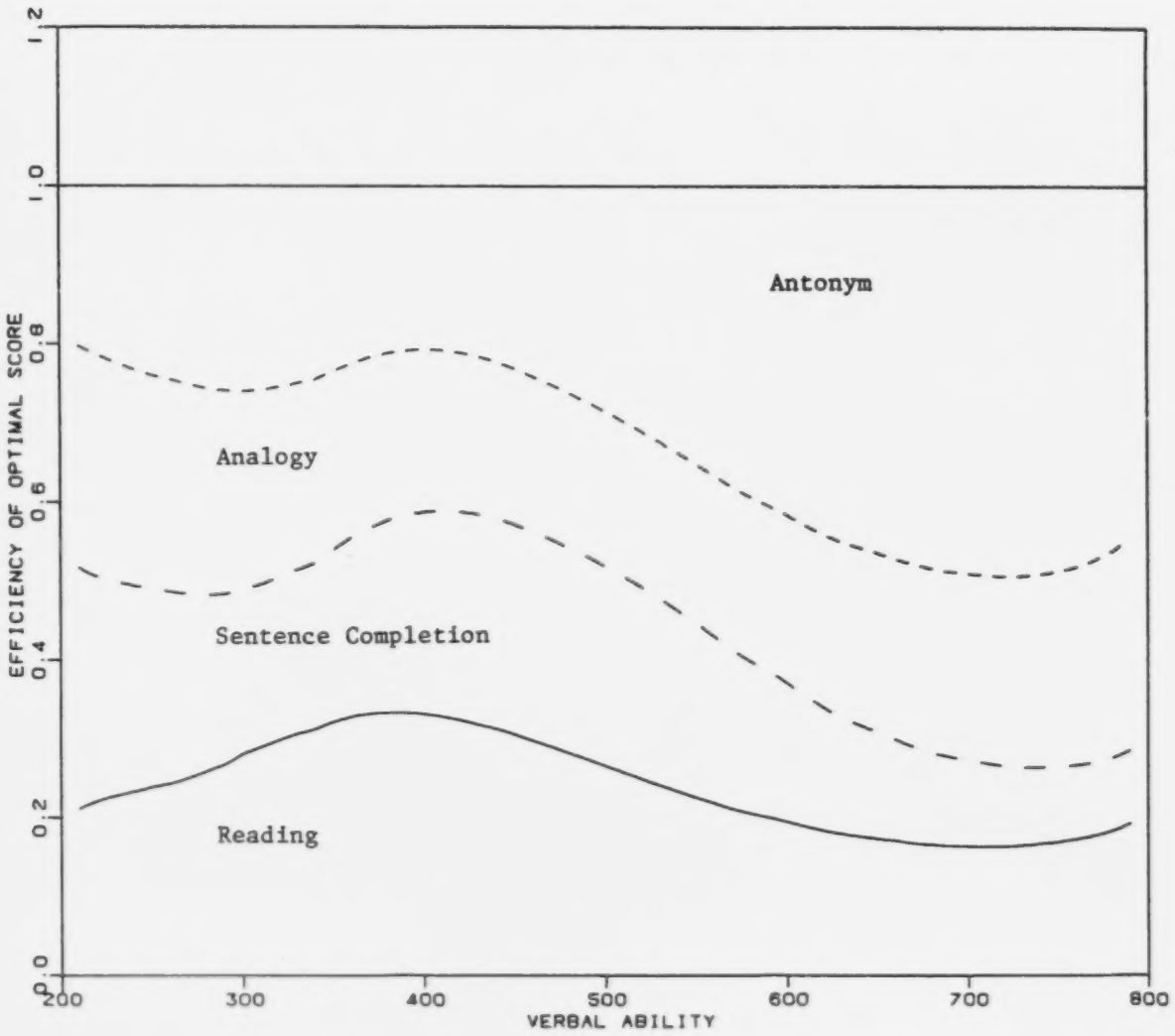

Figure 2. Partitioning of the measurement effectiveness of the optimally scored total verbal test into the additive contributions of the four item types.

Since the GRE is actually scored number-right, Figure 2 does not properly represent the actual scoring. Figure 3 is therefore presented, based on number-right scoring, but otherwise identical to Figure 2. Here the contributions of the item types do not add up to exactly 1.0, but they come close to doing so. Thus Figure 3 again shows that antonym items supply much of the measurement effectiveness at higher ability levels.

The large actual contribution of antonym and analogy items to effective measurement at high ability levels is probably due to the fact that most of the difficult items in the test are antonyms and analogies. What is the relative effectiveness of different item types when differences in item difficulty are removed?

Table 3 shows the relative efficiency distribution of the estimated item response theory item difficulty parameter b for each item type. To investigate the question of whether relative effectiveness differs by item type when differences in item difficulty are removed, subsets of items were selected so that the relative frequency distributions were matched. For sentence completion, a subset of 31 items was selected, solely by their b values, to have the same relative frequency distribution as the 143 reading items shown in Table 3. A subset of 43 antonym items and a subset of 49 analogy items were similarly selected, matched with the reading items on distribution of item difficulty (b).

Figure 4 shows the efficiency for measuring 'verbal ability', relative to the reading items, of number-right score on the remaining three item types when each type is matched with the reading items on distribution of item difficulty (b). The relative efficiency curves are adjusted for test length: They represent the results that would be obtained when the subtests compared are all of the same length.

The results show that, when matched on item difficulty, antonym items are more effective for measuring 'verbal ability' (as defined by the present GRE total verbal test) than reading items for examinees above 450 in true 'verbal ability'. The analogy items are better than reading items over the whole range of ability. The sentence completion items are better than reading items above 340.

This conclusion holds when the same length test is administered for each item type. Actually, reading items require more testing time than the other item types, so in practice fewer reading items could be administered in the available testing time, making the conclusion still more unfavorable to the reading items.

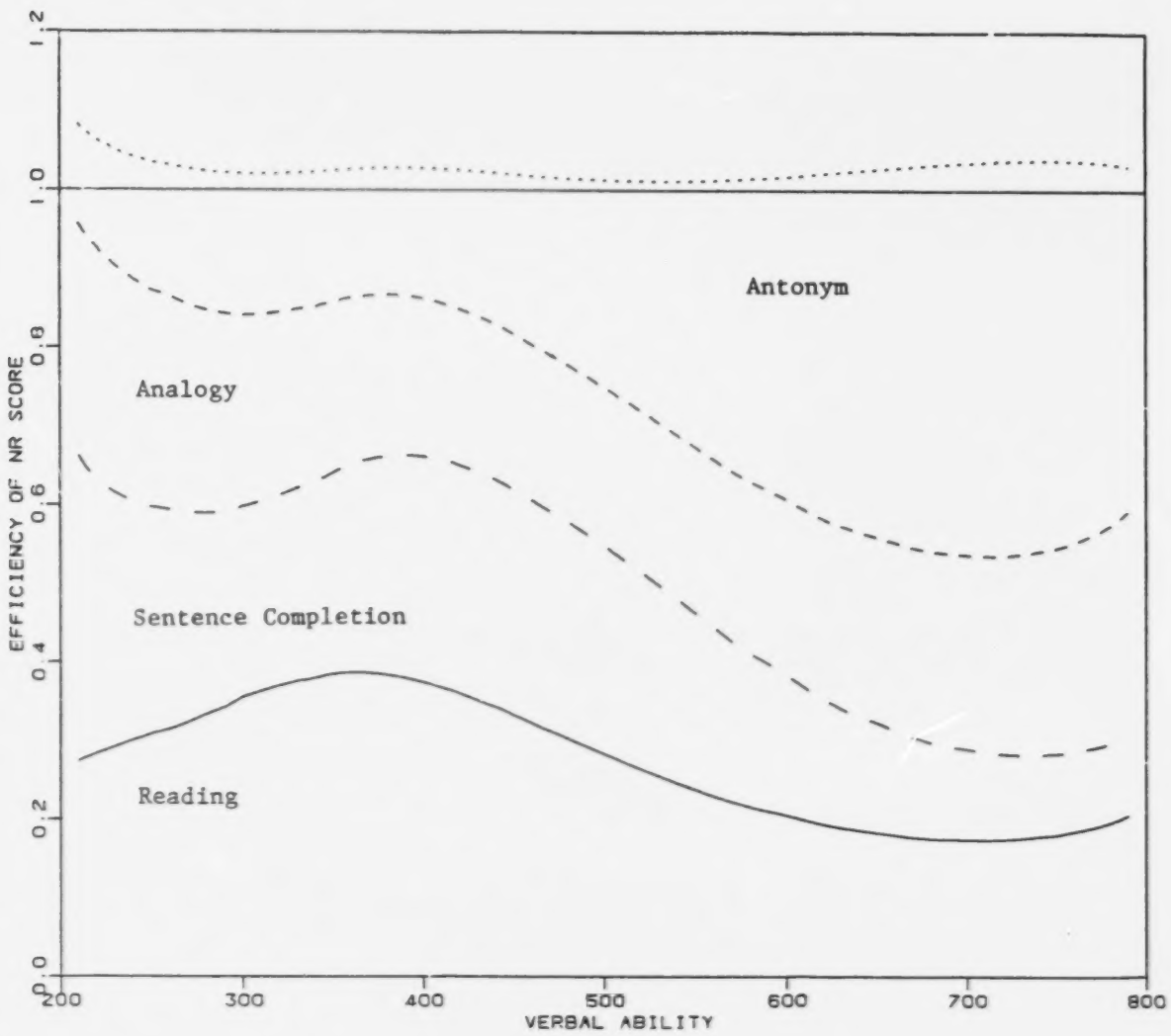

Figure 3. Approximate partitioning, as in Figure 2, for number-right scoring.

Table 3

Distribution of the Estimated Item Difficulty Parameter (b), Separately by Item Type

	Percent Within Item Type			
b	Reading Comprehension n = 143	Sentence Completion n = 91	Anology n = 117	Antonym n = 143
				0.7
3.0 -------				
			1.7	0.7
2.6 -------				
	1.4		2.6	3.5
2.2 -------				
	3.5	1.1	0.9	2.8
1.8 -------				
	2.1	2.2	17.1	15.4
1.4 -------				
	5.8	2.2	11.1	14.7
1.0 -------				
	12.6	5.5	12.8	14.7
.6 -------				
	8.4	16.5	7.7	9.1
.2 -------				
	12.6	15.4	8.5	5.6
-.2 -------				
	11.9	11.0	5.1	3.5
-.6 -------				
	12.6	13.2	4.3	6.3
-1.0 -------				
	14.7	3.3	5.1	3.5
-1.4 -------				
	6.3	2.2	4.3	3.5
-1.8 -------				
	2.1	4.4	5.1	5.6
-2.2 -------				
	1.4	4.4	3.4	5.6
-2.6 -------				
	3.5	6.6	6.8	1.4
-3.0 -------				
	1.4	12.1	3.4	3.5
	100.0	100.0	100.0	100.0

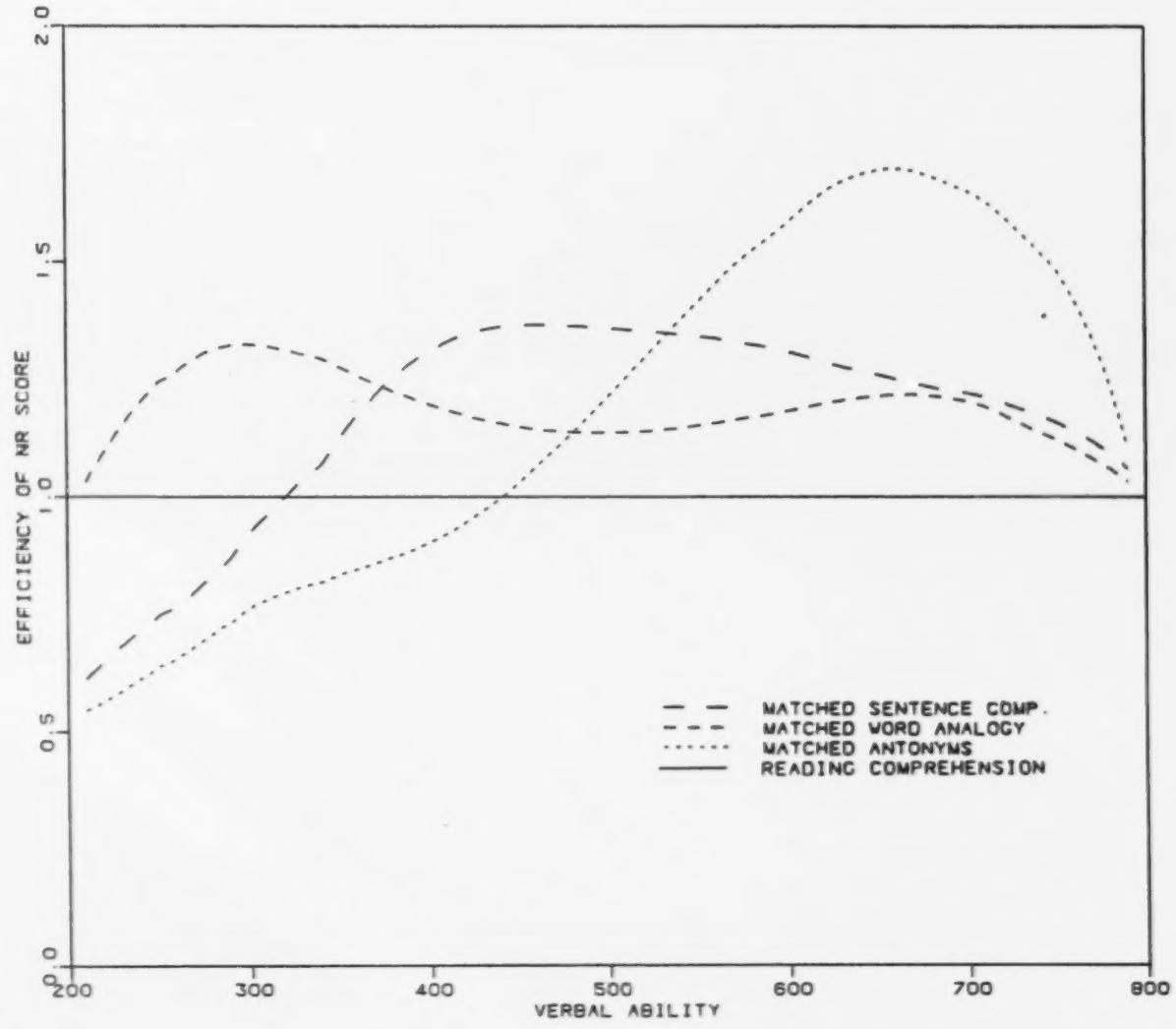

Figure 4. Efficiency, item for item of three item types, relative to reading items, for measuring 'verbal ability'; items are matched to reading items on difficulty.

7. Results for 'Reading Ability'

In this section, the criterion to be measured is 'reading ability' as defined by number-right true score on the GRE reading items. Separately for each item type, Figure 5 shows the efficiency for measuring 'reading ability', relative to the number-right score on the reading items. An adjustment is made so that the comparison is between subtests of equal length. Clearly the other item types, as represented in the GRE verbal test, are not as effective item-for-item for measuring 'reading ability.'

Figure 6 provides the same comparison as Figure 5, except that now each item type is matched on distribution of item difficulty (b) to the reading test. The conclusion is still similar: If one wishes to measure 'reading ability', item-for-item, this is best done with reading items. As would be expected from the item type correlations in Table 2, sentence completion items are the next most efficient measure of reading while antonyms are the least efficient measure of reading.

It does appear that other item types do better below 300 and above 750. This occurs because discriminating items, by definition, concentrate most of their effectiveness in a limited range of ability, whereas less discriminating items, by definition, spread their discrimination over a broader range. Since the nonreading item types are less discriminating for reading ability than most reading items, they necessarily are more effective than typical reading items outside the ability range where typical reading items discriminate well. The same effect would be observed if reading items with low discriminating power were substituted for the nonreading item types.

Figure 6 differs from Figure 4 primarily in the substitution of 'reading ability' for 'verbal ability' as the criterion to be measured. The fact that the figures are so different indicates that the two abilities differ substantially.

8. Summary and Discussion

The main purpose of this study is to compare the contribution of each of the four verbal item types to the measurement accuracy of the GRE verbal score. Sections 5 and 6 discuss accuracy for measuring the verbal ability defined by the GRE verbal score as it now exists (a composite of all four item types) while section 7 discusses accuracy for measuring reading ability defined by just the GRE reading comprehension items. In all three sections, the fact that reading comprehension items take more time than other types was ignored. In some cases corrections were made so as to compare subtests of the same 'length' as measured by the number of items in a subtest. In no case were adjustments made to compare tests requiring the same amount of administration time.

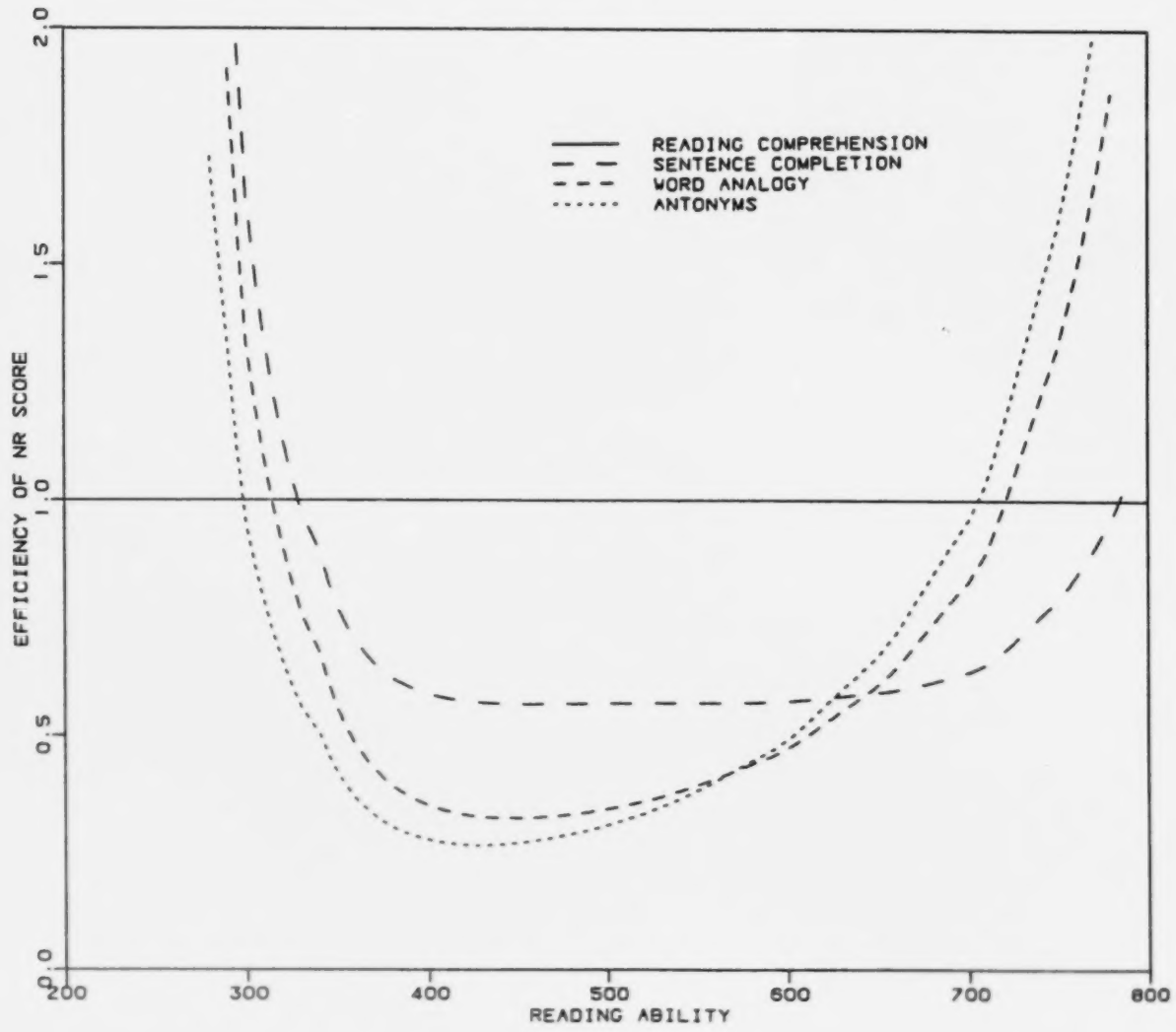

Figure 5. Efficiency, item for item, of three item types, relative to reading items, for measuring 'reading ability'.

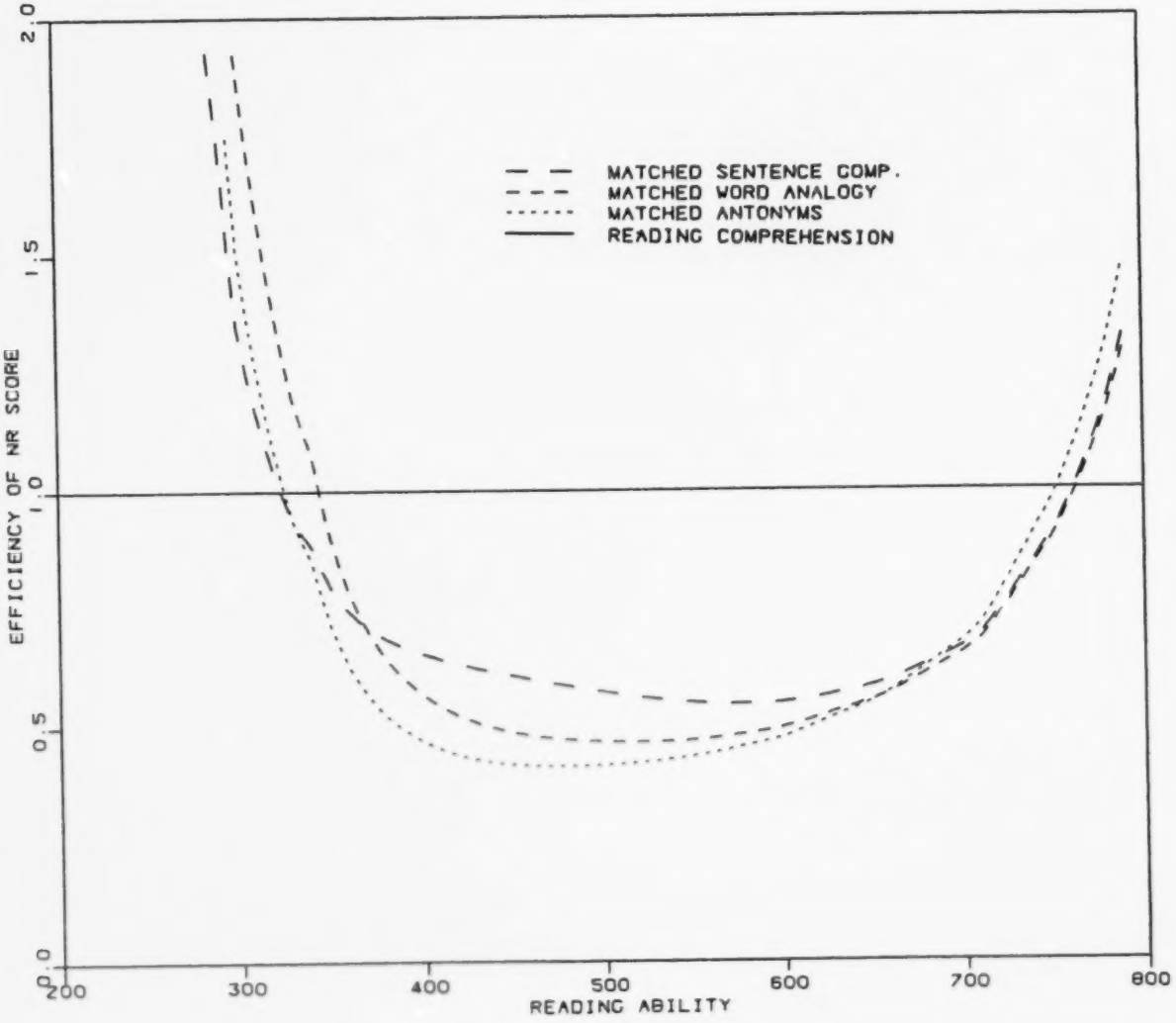

Figure 6. Efficiency, item for item, of three item types, relative to reading items, for measuring 'reading ability'. Items are matched to reading items on difficulty.

The testing time requirements by item type are summarized in Table 4. About one-and-a-half sentence completion items, three antonym items, and two analogy items can be taken in the same time required to answer one reading comprehension item. When testing time is taken into account, reading comprehension items appear less favorable than the results presented in the previous sections would suggest.

Table 4

Summary of Testing Time by Item Types

	Seconds per Item	Minutes to Take 76 Questions	Possible Number of Questions in 60 Minutes
Reading Comprehension	60	76	60
Sentence Completions	41	52	87
Antonyms	22	28	164
Analogies	27	34	133

For example, Figure 4 illustrates that when the item types are matched with reading comprehension items on distribution of item difficulty, antonym items are more effective for measuring 'verbal ability' than reading items for examinees above 450 in 'true ability'. This is true of tests of the same number of items. If you assume that 76 questions (the number in the GRE verbal measure) are administered, 28 minutes of testing time using antonyms is more effective than 76 minutes of testing using reading comprehension items for examinees with scores above 450. Even at a score of 200 (Figure 4) the relative efficiency of a single antonym item to a single reading comprehension item is above .5. This means it would take twice as many antonym items as reading comprehension items to measure 'verbal ability' with equal effectiveness. In terms of testing time, however, more than twice as many antonym items can be given as reading comprehension items in a given time period (see Table 4). Thus for matched difficulty distributions, it appears that, for a fixed time allotment, antonym items are as effective at measuring 'verbal ability' as reading comprehension items across the entire score range.

Item-for-item, the antonym items contribute the most accuracy for the higher ability examinees (above a true score of about 550). Analogy items provide a consistent contribution to accuracy all along the score scale. This is especially evident when the item types are matched on difficulty. Item-for-item, reading comprehension and sentence completion item types are

generally the largest contributors to accuracy of measurement of 'verbal ability' below a true score of about 450. At no point on the current 'verbal ability' scale is reading comprehension the most efficient item type when either the numbers of items or item difficulty distributions are matched.

Reading comprehension items are more efficient than the other item types for measuring the criterion of 'reading ability'. If you evaluate the results in terms of efficiency in testing time, rather than item-by-item, the differences in efficiency in measuring 'reading ability' mainly disappear. The minimum relative efficiencies, item-by-item, of antonyms, analogies, and sentence completions compared to reading comprehension are .4, .5, and .5, respectively. This means that, at the point on the scale where antonyms are measuring 'reading ability' least effectively, two-and-a-half times as many antonym items as reading items would be needed to measure 'reading ability' at least as well as reading comprehension throughout the scale. Since about three times as many antonym items as reading comprehension items can be given in the same time period (see Table 4), antonyms can be considered as effective as reading comprehension items in measuring 'reading ability', testing minute for testing minute. A similar argument can be made for analogy items. Sentence completion items would require slightly more testing time than reading comprehension to obtain the same efficiency at all points on the scale. However, since the results suggest that the four item types are not unidimensional, this conclusion may or may not be true.

What does this all suggest about the implications of the test content on validity issues at various score points? If one could assume that the four verbal item types are unidimensional (that is, they all measure the same thing), one could fairly strongly conclude that eliminating the analogy and antonym item types would seriously decrease the accuracy of higher 'verbal ability' scores (above about 600) and therefore potentially limit the validity of scores in this region. The evidence in section 7 of this report does suggest, however, that reading comprehension is measuring something different from what is being measured by the other verbal item types. Since the above conclusions depend on the assumption that all item types measure the same verbal ability, the evidence suggesting that the assumptions are only approximately met also implies that definite conclusions cannot be drawn.

The results do suggest, however, that caution be used in determining the appropriate content of the GRE verbal measure at upper score levels. Validity studies such as those described in Section 3 of the report would be necessary to obtain a clear resolution of this issue. In the absence of such definitive studies (such data are extremely difficult to obtain), it is important to consider a number of sources of data in determining test content. Continual review of the test specifications, including studies of both internal and external criteria are recommended as a way of assuring the continued validity of the GRE verbal measure.

References

Burton, N. W., & Turner, N. J. (1983). Effectiveness of the Graduate Record Examinations for predicting first year grades: 1981-82 summary report of the Graduate Record Examinations validity study service. Princeton, NJ: Educational Testing Service.

Kingston, N. M., & Dorans, N. J. (1982). The feasibility of using item response theory as a psychometric model for the GRE Aptitude Test (GRE Board Professional Report 79-12P). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Ramist, L. (1981). Validity of the SAT-verbal subscores. Internal memorandum. Princeton, NJ: Educational Testing Service.

Schrader, W. B. (1984). Three studies of SAT-verbal item types (College Board Report 84-7 and ETS RR 84-3). New York: College Entrance Examination Board.

Wilson, K. M. (1985). The relationship of GRE General Test item-type part scores to undergraduate grades (GRE Board Professional Report 81-22P). Princeton, NJ: Educational Testing Service.

Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), Application of item response theory. Vancouver, B.C.: Educational Research Institute of British Columbia.